

## **Civil Rights Data Collection: State and National Estimation Data Notes**

### **About the CRDC Data**

The CRDC has generally been collected biennially from public school districts in each of the 50 states and the District of Columbia. Data are collected for each school in the districts included in the survey. For the first time, the 2009-10 CRDC was collected in two parts. Part 1 is “snapshot” data related to enrollment and Part 2 is cumulative and “end-of-year” results” data. The 2009-10 CRDC contains information from a sample of about 7,000 school districts and over 72,000 schools in those districts.

The Office for Civil Rights (OCR) at the U.S. Department of Education strives to ensure CRDC data is an accurate and comprehensive depiction of student access to educational opportunities in sampled school districts. The submission system includes a series of embedded edit checks to ensure significant data errors are corrected before the district submits its data. Additionally, each district is required to certify the accuracy of its submission. Only a district superintendent, or the superintendent’s designee, may certify the CRDC submission. Ultimately, the quality of the CRDC data depends on accurate collection and reporting by the participating districts.

### **About the State and National Estimations**

Then national and state estimations presented in this report are based on data reported by a sample of public school districts. Data from school districts is self-reported.

The sample estimates may differ somewhat from the values that would be obtained from a universe of respondents. The difference occurs because a sample survey estimate is subject to two types of errors: nonsampling and sampling. Nonsampling errors are attributed to many sources, including definitional difficulties, the inability of respondents to provide accurate data, differences in the interpretation of questions, errors made in collection (e.g., in recording or coding the data), and errors made in estimating values for missing data. Quality control and edit procedures were used to reduce errors. However, users should carefully consider the caveats for analyzing the state and national estimations included in this document.

### **Sample Design**

The CRDC has used a stratified random sample to ensure that the sample of schools contains sufficient numbers for reliable estimates. Some districts are always in the sample, such as all districts in states with fewer than 25 public school districts, and districts subject to federal court orders that are monitored by the U.S. Department of Justice. The remaining districts have

been chosen using a rolling stratified sampling method that ensures that a representative group from each state is included in the data collection. For the 2009-10 CRDC all districts with more than 3,000 students, state-operated facilities for students who are deaf or blind, and state-operated long-term secure juvenile justice facilities were also included in the CRDC sample.

The CRDC sample for the 2009-10 CRDC was selected from a universe list of public school districts compiled and maintained by the National Center for Education Statistics (NCES) in the Common Core of Data (CCD) Universe of Public School Districts. The CCD includes regular and nonregular schools (special education, alternative, vocational, or technical), public charter schools, and Bureau of Indian Education schools. The sampling frame was adjusted from the CCD in order to fit the definition of a school eligible for CRDC. The CCD includes some schools that do not offer classroom instruction in grades PK-12 or the ungraded equivalent. In some instances, schools in the CCD are essentially administrative units that may oversee entities that provide classroom instruction or they may only provide funding and oversight. Additional school records were added to the sampling frame. Most of these records were for alternative, special education, or long-term secure juvenile justice facilities included in data maintained by the Department of Justice. Following the initial sample selection, to broaden the coverage of the CRDC, the sample was augmented to include all school districts with more than 3,000 students using the 2008-09 Universe of Public School Districts.

### **Sample Weights**

Weighting of the sample school districts was carried out to produce national and state estimates. The weighting procedures used in the CRDC are designed to take into account the district's selection probability; to reduce biases that may result from nonresponse; and to make use of available information from external sources to improve the precision of sample estimates.

The general purpose of weighting is to scale up the sample estimates to represent the target survey population. For the CRDC, a base weight (e.g., the inverse of the school district's probability of selection) is used as the starting point. Next, a series of adjustment factors were calculated and applied using information from the 2009-10 Universe of Public Schools to ensure student count estimates aligned, to the extent possible, with state and national membership counts of students in CRDC eligible schools.

### **Response Rates**

Traditionally the CRDC has achieved very high overall response rates. Every state, with the exception of Minnesota, had an overall response rate of 100%. One district in Minnesota provided partial data for Part 2 and as a result, the overall response rate for Minnesota dropped to 99.6%.

To provide users with a way of gauging the amount of missing data used in the statistical projections, item-specific response rates were computed for each survey item. Response rates were generated based on the 1,250 variables estimated for each state and the nation. An important consideration for response rates is that the reporting process for 2009-10 required all school and districts respond to each question on the CRDC prior to certification. Null or missing data prevented a school district from completing their CRDC submission to the Office for Civil Rights. Therefore, in cases where a school district may not have complete data, some schools or districts may have reported a zero value in place of a null value. It is not possible to determine all possible situations where this may have occurred. As such, it may be the case that the item response rates may be positively biased.

The response rate for each particular item is computed as  $R=100(c/(c+n))$  where  $c$  is the number of schools reporting a non-negative (zero or greater) value for the given item and  $n$  is the number of schools for which the given item should have been reported but is missing. Note that nonapplicable items are excluded from the calculations. For example, an elementary school with grades kindergarten through sixth grade would be excluded in the calculating item response rates for grade 12 retention.

### **Race/Ethnicity Categories**

Users will note that for the 2009-10 CRDC, districts had the option of reporting one of two different sets of race and ethnicity categories. All school districts were able to choose to report data by the traditional five race and ethnicity categories (Hispanic, White, Black/African-American, Asian/Pacific Islander, and American Indian/Alaska Native). Districts that met all of the requirements of the Department's 2007 Final Guidance on Collecting, Maintaining and Reporting Data on Race and Ethnicity had the option of reporting data using the seven race and ethnicity categories (Hispanic/Latino, White, Black/African-American, Asian, Native Hawaiian/Other Pacific Islander, American Indian/Alaska Native, and Two or More Races). Every school within a district must report using the same set of five or seven race and ethnicity categories. For more information on the Department's 2007 guidance, please visit <http://www2.ed.gov/policy/rschstat/guid/raceethnicity/index.html>.

Less than one quarter of school districts reported data under the seven race and ethnicity categories. Given the small number of school districts reporting under the seven race and ethnicity categories, data reported under the seven categories was mapped back into the five race/ethnicity categories. To map data from the seven race and ethnicity categories into the five, data from the Native Hawaiian and Pacific Islander category was combined with data from the Asian category to form a single Asian/Pacific Islander category. Data provided in the category of two or more races was

allocated to the five race and ethnicity categories by examining the proportion of American Indian, Asian/Pacific Islander, Black, Hispanic, and White students from the Common Core of Data.

## **Imputations**

Several new items were added to the CRDC in the 2009-10 school year. Because this was the first year of collection for many data elements, some school districts were unable to provide complete and accurate data for these new data items or may have made reporting errors such as reporting the number of incidents of discipline as opposed to the number of students disciplined.

To account for outliers within the data set that may be the result of reporting errors, an imputations process was employed. The process for imputing values was completed through a series of steps designed to minimize the impact of reported values identified as outliers on the state and national estimations. Values identified as “too small” or “too large” were adjusted toward the median. For example, based on the school district reported data, only 5 districts account for over 50 percent of the non-imputed national estimate for teacher. These same districts represent less than 0.6% of total enrollment. To account for these types of errors, the imputations process replaced improbable values in the dataset, such as decimal placement errors resulting in thousands of teachers per students.

Several methods for imputation were evaluated. Initially, and consistent with more traditional methods, the imputation process was applied after completion of the projections, with outliers or more problematic values identified at this point in time. However, in several iterations of generating projections and standard errors using this imputation process, outliers in the existing CRDC school data were impacting values for standard errors.

Imputations were therefore applied prior to generating the projections and standard errors to remove issues associated with extreme outlier values. An imputations process modeled on the winsorizing models commonly used in nonparametric procedures where outliers are pulled to the median to mitigate their effect was ultimately decided upon. Consistent with methodology employed by other U.S. Department of Education survey datasets, the total enrollment of the school was indexed to determine the relative magnitude of a variable as an outlier or extreme value. Further, the plausibility of ‘0’ being a median value was also evaluated and corrections provided to adjust for this situation.

While this imputation process mitigated the influence of outliers on the state and national estimations by adjusting values towards the median, the imputations process may not resolve widespread underreporting or data quality issues.

## **Standard Errors**

Estimates derived from a probability sample are subject to sampling error because only a fraction of the target population has been surveyed. Each estimate will have an associated standard error, which provides important information pertaining to the accuracy of the calculated estimate. The standard error provides important information about the reliability of a sample-based estimate because it reflects the degree of uncertainty that results from the use of a sample rather than a complete “census” of the survey population. In particular, the standard error can be used to construct intervals within which the “true” population value can be expected to lie with a high level of confidence.

Estimates with large statistical uncertainty are marked with a single asterisk (\*). In these cases, the standard error divided by the estimate is greater than 0.30. Estimations where the standard error exceeds 50% of the estimate have been suppressed.

## **Comparison to Existing Data Sources**

For some state and national estimations, the U.S. Department of Education has published comparable data based on a similar universe of school districts from which the CRDC drew its sample. Estimates which vary from published comparable data sources are noted with a single dagger (†). Users should exercise caution when analyzing such cases, as the CRDC estimate may be effected by possible district reporting errors.

## **Cautions for Analyzing State and National Estimates**

Care must be taken in comparing estimations for various student populations because some of the difference may not be attributable to the education system. Some of the difference may be due to sampling or non sampling error.

OCR strongly encourages users when making comparisons to test the selected findings for statistical significance using Student’s *t*-test to ensure that the differences are larger than those that might be expected due to sampling variation. Users should also carefully consider comparisons among student groups where the membership total has been flagged with a single dagger (†) as it is possible that underlying data quality issues may be effecting the estimation for the student group.

Estimations were calculated separately for the total counts and the individual disaggregations. As a result, users may also find that the sum of the disaggregated counts may not always equal the reported totals.

## **Data Considerations and Anomalies**

This section provides some caveats and considerations that users should take into account when using the state and national estimations. The following are caveats users should carefully consider when analyzing the new CRDC data items.

- Users should exercise caution when comparing the Part 1 number of students enrolled in Algebra 1 to the Part 2 number of students passing Algebra 1. The Part 1 data on the number of students enrolled in Algebra I was collected from the Fall of 2009, while the number of students passing Algebra I was collected at the end of the 2009-10 school year. Due to changes in enrollments over the school year, the count for enrollment and passing may not cover the exact same set of students.
- The 2009-10 CRDC expanded the data collected on various disciplinary actions. Users should exercise caution in analyzing outliers in the data set since districts may have accidentally reported the number of disciplinary incidents versus students disciplined. Additionally, some districts, unable to report complete and accurate data for school-related arrests and referrals to law enforcement, may have reported zero students in these categories. As a result, these data may be underestimated.
- The 2009-10 CRDC collection included the first of its kind school and district level reporting of bullying and harassment. However, many districts were unable to provide complete data on students reported to have been bullied or harassed, students disciplined, and incidents by type of bullying and harassment. As a result, these data may be underestimated.

### **Contact information**

If you have any comments or questions concerning the use of CRDC data, please write to:

CRDC Survey Coordinator  
Office for Civil Rights  
U.S. Department of Education  
400 Maryland Avenue S.W.  
Room 4E318  
Washington, D.C. 20202-1172

Email: [ocrdata@ed.gov](mailto:ocrdata@ed.gov)